TITLE OF THE INVENTION

MEMORY SYSTEM FOR IMPROVING DATA INPUT/OUTPUT PERFORMANCE AND METHOD OF CACHING DATA RECOVERY INFORMATION

CLAIM ~~FOR~~ OF PRIORITY

This application makes reference to, ~~incorporates the same herein~~, and claims all benefits accruing under 35 U.S.C. §119 from an application for *MEMORY SYSTEM FOR IMPROVING DATA INPUT/OUTPUT PERFORMANCE AND METHOD OF CACHING DATA RECOVERY INFORMATION* earlier filed in the Korean Industrial Property Office on the 16th of September 1996, and there duly assigned Serial No. 40202/1996~~, a copy of which application is annexed hereto~~.

BACKGROUND OF THE INVENTION

~~Technical Field~~ Field of the Invention

The present invention relates to a memory system such as ~~redundant arrays~~ a redundant array of inexpensive disks (RAID) and, more particularly, to a redundant ~~arrays~~ array of inexpensive disks capable of providing high ~~performance of data input/output operation~~ data input/output performance and a method of caching data recovery information using the redundant ~~arrays~~ array of inexpensive disks.

~~Related Art~~ Description of the Related Art

~~High~~ A high-technology computer system depends considerably on its central processor unit (CPU)

and input/output subsystem to increase overall system performance. While information processing speed of the CPU has been dramatically improved in recent years because of VLSI technology, the performance of the input/output subsystem has not improved as desired. This increases the time required to access data in the memory system. Furthermore, as the ~~cost needed to restore data is~~ increased when an error is generated in the input/output subsystem, an input/output subsystem having excellent performance and reliability is needed. As a solution to this, a disk array system known as a redundant array of inexpensive disks (RAID), constructed of a number of relatively small capacity disk drives has been proposed as a low cost alternative to a single large expensive disk for ~~providing large storage of~~ digital information.

RAID systems are now commercially available as cost effective mass storage providing reliable and continuous services to a host computer or network file server. The theory of RAID is to use relatively inexpensive disks, which may individually have a higher chance of failure than expensive disks, and compensating for this higher failure rate by adding redundancy by creating and storing parity blocks to facilitate recovery from a disk failure. Reports on the performance and reliability of disk arrays are presented in "*A Case For Redundant Arrays Of Inexpensive Disks (RAID)*" by D. Patterson, G. Gibson, and R. H. Kartz, at Report No. UCB/CSD 87/89, December 1987, Computer Science Division (EECS), University of California, Berkeley, Calif. 94720. Exemplars of contemporary RAID systems are disclosed in U.S. Patent No. 5,257,367 *for Data Storage System With Asynchronous Host Operating System Communication Link* issued to Goodlander et al., U.S. Patent Nos. 5,367,669 and 5,455,934 for *Fault Tolerant Hard Disk Array*

Controller issued to Holland et al., U.S. Patent No. 5,418,921 for *Method And Means For Fast Writing Data To LRU Cached Based DASD Arrays Under Drivers Fault Tolerant Modes* issued to Cortney et al., U.S. Patent No. 5,463,765 for *Disk Array System, Data Writing Method Thereof, And Fault Recovering Method* issued Kakuta et al., U.S. Patent No. 5,485,598 for *Redundant Disk Array (RAID) System Utilizing Separate Cache Memories For The Host System And The Check Data* issued to Kashima et al., U.S. Patent No. 5,522,032 for *RAID Level 5 With Free Blocks Parity Cache* issued to Franaszek et al., U.S. Patent No. 5,530,948 for *System And Method For Command Queuing On RAID Levels 4 And 5 Parity Drives* issued to Islam, U.S. Patent No. 5,579,474 for *Disk Array System And Its Control Method* issued to Kakuta et al., U.S. Patent No. 5,640,506 for *Integrity Protection For Parity Calculation For RAID Parity Cache* issued to Duffy, and U.S. Patent No. 5,636,359 for *Performance Enhancement System And Method For A Hierarchical Data Cache Using A RAID Parity Scheme* issued to Beardsley et al.

As generally discussed in the Patterson report and subsequent contemporary RAID systems ~~as set forth~~, the large personal computer market has supported the development of inexpensive disk drives having a better ratio of performance to cost than single large expensive disk systems. The number of input/outputs (I/Os) per second per read/write head in an inexpensive disk is within a factor of two of the large disks. Therefore, the parallel transfer from several inexpensive disks in a RAID system, in which a set of inexpensive disks function as a single logical disk drive, produces ~~greater~~ better performance than a single large expensive disk (SLED) at a reduced ~~price~~ cost.

Unfortunately, when data is stored on more than one disk, the mean time to failure varies inversely with the number of disks in the array. In order to correct for this decreased mean time to failure of the system, error recognition and correction is a characteristic of all RAID systems. ~~Generally, each RAID system is organized in six structures commonly referred to as six levels each,~~ having a different means for error recognition and correction as described hereinbelow.

In a RAID structure of level 0, data is distributed and stored in all drives in the disk array, taking interests in performance rather than data reliability.

In a RAID structure of level 1, the mirroring, a conventional method of improving the disk performance, ~~needs a lot of costs~~ has a high cost since all contents of the disk must be stored in a reproduction disk without change. Accordingly, in a database system requiring a large-capacity disk space, only ~~the~~ fifty ~~percents~~ percent of the disk space can be used. However, ~~the~~ mirroring is the best way to enhance the data reliability because identical data is stored in the reproduction disk. In a RAID structure of level 2, this is used to minimize the cost required to enhance data reliability. The RAID structure of level 2 distributes and stores data in each disk array in ~~bite,~~ bites and has several test disks using a ~~hamming~~ Hamming ~~code~~ Code, besides the data disk, in order to recognize and correct errors.

In a RAID structure of level 3, data is input/output in parallel to/from the drive when input/output is requested once, and parity data is stored in a separate drive. Furthermore, disk spindles are synchronized so as to make all drives simultaneously input or output data. Accordingly,

rapid data transmission can be carried out even if parallel input/output is not performed fast. If one drive has error, the erroneous data can be restored by using the currently operated drive and parity drive even though the total data rate is decreased. The RAID structure of level 3 is used in an ~~application which requires very fast data transmission rate, super computer and image manipulation~~ processors. That is, the RAID of level 3 has ~~higher~~ efficiency in a long data block transmission but has lower efficiency in a short data block transmission which requires fast input/output request. Furthermore, since the data drive is used together with a single drive for redundancy, ~~the drive~~ smaller than that used in the RAID of level 1 is used but its controller becomes more expensive and complicated.

In RAID structure of level 4, the parity data is calculated and stored in a separate drive, and data is striped across. The data can be restored when it has error. Its reading performance is similar to that of RAID of level 1 but its writing is much poorer than the single drive because the parity information must be provided to the single drive. Thus, the RAID structure of level 5 having improved writing performance is supplemented to the RAID of level 4.

In RAID structure of level 5, data is striped across in each drive array, and parity data is distributed and stored in all drives in order to remove bottleneck phenomenon when data is written. In this RAID structure, since the data written in all drives must be read in order to calculate the parity when the data is written, its speed is slower. However, it is possible to process the data input/output transmission and to restore data stored in a drive having error. Accordingly, the RAID structure of

~~level 5 is effective in recording of long data, and also effective in recording of short data if an~~ application program gives weight ~~on~~ the data reading or the array design is improved in order to increase the writing performance. Even if the size of the data block is decreased, performance and data availability can be obtained to some degree. Moreover, the RAID structure of level 5 is most effective in terms of cost in comparison with a non-array device.

Among all disk array structures, the RAID structure of level 5 provides a higher reliability with smaller additional cost, and at the same time, makes the parallel disk access possible, resulting in the improvement of data processing rate. Generally, when data writing instruction is received from the host computer for writing in each drive in the RAID structure of level 5, the CPU determines a target location, and transmits the data to ~~controller 6 where~~ old data and old parity stored in each drive are read. ~~Controller~~ calculates a new parity based on an exclusive OR arithmetic operation, and writes new data and new parity in a predetermined drive. However, when writing instruction of a short data block is received from the host computer in the RAID structure of 5 level, access of another disk on the strip is brought about which ~~attributes~~ to a deterioration of the entire system performance. I have observed that this phenomenon ~~remarkably~~ appears in the on-line transaction processing environment having many operation loads. That is, in case of the partial strip writing, old parity and old data are read from a predetermined drive, exclusive-OR operation is performed to determine new data, and then new parity information and new data are written in the predetermined drive. Two-time reading and writing operations are necessarily required which results in a larger overhead of data write in comparison with a single large expensive drive.

SUMMARY OF THE INVENTION

Accordingly, it is therefore an object of the present invention to provide a redundant ~~arrays~~ array of inexpensive disks (RAID) system with an enhanced process performance and a reduced overhead of data write.

It is also an object to provide a RAID system capable of reducing an overhead during a read operation of data recovery information in order to improve its data input/output performance, and a method of caching data recovery information using the memory system.

These and other objects of the present invention can be achieved by a redundant ~~arrays~~ array of inexpensive disks (RAID) system which includes a plurality of defect-adaptive memory devices for sequentially storing information needed for data recovery in a predetermined region of a recording medium in the form of block, and storing data in a region other than the predetermined region. A plurality of caches are connected to the adaptive memory devices to store information blocks needed for data recovery, the information blocks being read from a predetermined memory device. A controller is connected to each adaptive memory device and cache to control the writing and reading of data and information needed for data recovery in each memory device, calculate information needed for recovery of data read from each memory device, and store the information needed for recovery of data calculated in a predetermined cache.

The present invention is more specifically described in the following paragraphs by reference to the drawings attached only by way of example.

BRIEF DESCRIPTION OF THE DRAWINGS

A more complete appreciation of the present invention, and many of the attendant advantages thereof, will become readily apparent as the same becomes better understood by reference to the following detailed description when considered in conjunction with the accompanying drawings in which like reference symbols indicate the same or similar components, wherein:

FIG. 1 is a block diagram of a RAID system;

FIG. 2 illustrates an example of data transmission of the RAID system;

FIG. 3 is a flowchart illustrating a process of writing data and parity information transmitted from a host computer to each drive in the RAID system;

FIG. 4 is a block diagram of an RAID system constructed according to an embodiment of the present invention; and

FIG. 5 is a flowchart illustrating a process of writing data and parity information in the RAID system constructed according to the embodiment of the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Referring now to the drawings and particularly to FIG. 1, which illustrates a redundant ~~arrays~~ array of inexpensive disks (RAID) system in level 5. As shown in FIG. 1, the RAID system includes a central processing unit (CPU) 2, a controller 6 connected to the CPU 2 via an input/output bus 4, and a plurality of disk drives DR1-DR5 connected to the controller 6 via SCSI bus 8.

CPU 2 transmits data transmitted through an input/output bus 4 from a host computer (not

shown) to the controller 6. The controller 6 connected to input/output bus 4 is controlled by CPU 2 to control input/output data between drive disks DR1 to DR5 which are connected to CPU 2 and SCSI bus 8. Each drive DR1 to DR5 connected to SCSI bus 8 records and reproduces the data transmitted from the host computer under the control of controller 6.

FIG. 2 illustrates an example of data transmission of the RAID structure in level 5. Data ND transmitted from the host computer is divided by strip (the data is divided by strip 3 in FIG. 2), distributed and stored in each drive DR1 to DR5. That is, each drive DR1 to DR5 has a data block D in which data is stored, and a parity block P in which parity information is stored, to thereby store the data transmitted from the host computer under the control of controller 6.

FIG. 3 is ~~a control flow chart~~ a flowchart for explaining the writing of the data and parity information transmitted from the host computer in each drive in the RAID structure of level 5. Referring to FIG. 3, when a data writing instruction is received from the host computer, the CPU 2 calculates a target location at step 10. At step 12, CPU 2 transmits the data transmitted from the host computer to controller 6. Controller 6 reads old data OD and old parity OP stored in each drive at steps 14 and 16. Next, the controller 6 calculates a new parity NP according to the following formula (1).

NP=OP⊻OD⊻ND (⊻ means exclusive OR) ----- (1)

Controller 6 writes data ND and new parity NP in a predetermined drive at steps 20 and 22.

As described, ~~in case that~~ when a writing instruction of a short data block is received from the host computer in the RAID system of level 5 structure, access of another disk on the strip is brought about which ~~attributes~~ leads to a deterioration to the entire system performance. This ~~remarkably~~ appears in the on-line transaction processing environment having many operation loads. That is, in case of the partial strip writing, old parity OP and old data OD are read from a predetermined drive, exclusive-ORed according to formula (1), its result is exclusive-ORed with data ND, and then new parity NP and new data ND are written in a predetermined drive. Thus, two-time reading and writing operations are needed which results in a larger overhead of write data in comparison with a single large expensive drive.

Turning now to FIG. 4 which illustrates a RAID system to which parity cache arrays 38 are connected according to an embodiment of the present invention. Referring to FIG. 4, the RAID system consists of a CPU 30 for controlling the overall system. A controller 34 ~~which~~ is connected to CPU 30 through an input/output bus 32 to distribute and store data transmitted from a host computer to each drive array 39, or reproduce the stored data under the control of ~~CPU 3~~ CPU 30. Drives 1 to 5 (39) ~~which~~ are connected to controller 34 through SCSI bus 36 to store and reproduce the data and data recovery information (parity information) transmitted from the host computer under the control of controller 34. Caches 1 to 5 (38) ~~which~~ are connected to controller 34 and input/output bus 36 placed between drives 39 to store the parity information.

Each drive 39 consists of a plurality of blocks in order to store and read the data and parity

~~information. Furthermore, each drive 39 sets up the predetermined number of parity block from the~~ cylinder zero on the disk, and uses it as a parity information storing region, without using the stripping method defined in the RAID structure in level 5. Here, the data cannot be recorded in the parity information storing region.

FIG. 5 ~~a control flow chart~~ for explaining a process of writing data and parity information in the RAID system constructed according to the embodiment of the present invention. The control process of writing data will be explained in detail with reference to FIGs. 4 and 5 hereinbelow.

First of all, the data writing instruction is received from the host computer, CPU 30 updates a task file ~~required at~~ step 40, and then calculates a target cylinder (=parity block + request cylinder) in order to use a separate parity block in the drive. Then, CPU 30 transmits new data ND to be written ~~at~~ step 42. ~~Controller~~ 34 next reads old data OD from a predetermined drive 39 in order to generate new parity NP, and then examines if old parity information OP to be read is hit in cache 38 ~~at~~ step 46. Here, if the old parity information OP is hit in cache 38, controller 34 proceeds to step 50. If the old parity information is not hit in cache 38, controller 34 proceeds to step 48. That is, ~~in case that the old parity information OP and parity information are not hit, controller 34 reads the~~ old parity information OP from the predetermined drive ~~at~~ step 48, updates a cache table, and then moves to step 50. ~~Controller~~ 34 calculates a new parity NP by exclusive-ORing the old parity information read and the new data ND through the following formula (2).

NP=OP⊻OD⊻ND ---------- (2)

the controller loads and predetermined cache 38

~~Controller~~ 34 updates the cache table ~~at~~ step 52, and then writes the new data ND transmitted from the host computer and the calculated new parity NP in a predetermined drive 39 ~~at~~ steps 54 and 56. Then, the data writing process of the present invention is completed.

According to the present invention, the parity cache is connected between each drive and ~~controller in order to rapidly apply parity information read request. Furthermore, since the parity~~ block for storing the parity information is set up from the cylinder zero on the disk, it is now possible to prevent time delay due to a separate search when sequential read/write operation is carried out.

While there have been illustrated and described what are considered to be preferred embodiments of the present invention, it will be understood by those skilled in the art that various changes and modifications may be made, and equivalents may be substituted for elements thereof without departing from the true scope of the present invention. In addition, many modifications may be made to adapt a particular situation to the teaching of the present invention without departing from the central scope thereof. Therefore, it is intended that the present invention not be limited to the particular embodiment disclosed as the best mode contemplated for carrying out the present invention, but that the present invention includes all embodiments falling within the scope of the appended claims.